New Advances in Computer Modeling of Chemical and Biochemical Data

5764

Nonlinear regression analysis has been successful for quantitative modeling of many types of experimental data. Recent solutions to several long-standing problems in nonlinear regression are reviewed in this paper. These include removal of correlations among parameters as well as enhancing resolution in deconvolution of ill-resolved signals with many component parts.

Orthogonalization of parameter space can remove correlations, resulting in precise estimation of all parameters. Fourier deconvolution algorithms can be combined with nonlinear regression to resolve severely-overlapped component bands in protein fourier transform infrared (FT-IR) spectra, leading to complete secondary structural analysis. These examples show that in instances where nonlinear regression cannot stand alone, its combination with auxiliary mathematical methods can result in powerful analytical tools.

The ongoing microelectronics revolution makes ever increasing computer power routinely available to the physical scientist. Nonlinear regression analysis has proven extremely useful in situations where instrumental data must be analyzed quantitatively with mathematical models to obtain the information required by the experimentalist.

Recent application highlights of nonlinear regression in chemistry include titrations without the necessity of standardization, determination of large bimolecular rate constants from voltammetric data, multicomponent chemical analysis from time-resolved fluorescence spectra, automated mechanistic analysis, and extraction of diffusion coefficients for micellar aggregates from electrochemical data.[1] In the biochemical arena, new applications include determination of the molecular basis for salt induced solubility and colloidal stability of proteins, quantitative analysis of field dispersion of NMR relaxation rates for water interactions with proteins, and mechanisms of the time dependence of bacterial growth under a variety of environmental conditions.[2]

Some traditional problems have long plagued certain applications of nonlinear regression analysis. These include correlations among parameters and lack of methods for dealing with inherently ill-resolved signals with large numbers of component parts. The latter problem may also involve parameters which are highly correlated. In this paper, we review recent approaches to dealing with these two long-standing barriers.

First, we discuss removal of parameter correlation by orthogonalization of the parameter space. Second, we present a way to enhance resolution of badly overlapped bands from protein IR spectra via Fourier deconvolution prior to nonlinear regression. This approach has been applied to determining global secondary structure of proteins. Both methods are general and can, in

principle, be applied to a wide variety of types of experimental data.

I. Removal of Parameter Correlation by Orthogonalization

Correlation. Nonlinear regression analysis seeks to find the best set of parameters of a mathematical model with respect to a given set of data. With a general regression program,[1,2] the user writes code for the model into a subroutine, and provides the program with experimental data and starting "best guesses" for the parameters. The program then seeks a minimum in an appropriate error sum, usually by making use of the principle of least squares. By this principle, the best values of the parameters are found at the minimum of the error sum S (Eq. (1)):

$$S = \sum_{j=1}^{n} [y_j(\text{meas}) - y_j(\text{calc})]^2 \quad (1)$$

where y_j(meas) are experimental data and y_j(calc) are computed from the regression model for the n data points. The purpose of the model subroutine is to supply values of y_j(calc) to the main program.

S is minimized by systematic iterative variation of the parameters by an appropriate algorithm. The approach to minimum S starting from a set of initial "guesses" for m parameters can be thought of as the journey of the initial point toward the global minimum of an error surface in an m + 1 dimensional coordinate system called "parameter space." One axis of parameter space corresponds to S; the others correspond to the parameters. The minimum S is called the convergence point.

Two interdependent parameters in a regression model are said to be correlated. Their final values are interdependant and sensitive to the starting point of the computation. If correlation is strong, convergence and/or unique estimates of each parameter can be difficult to achieve. Parameter correlation is equivalent to having nonorthogonal axes in the parameter space.

An Example of Orthogonalization. Troublesome parameter correlation arises in obtaining chemical rate constants using a technique called chronocoulometry. In this method, a step of potential is applied to a working electrode and the accumulated charge passing through the electrochemical cell is recorded vs time. Models describing charge response in chronocoulometry are expressed in closed form as explicit functions of time. Below, we present the application of Gram-Schmidt orthogonalization to chronocoulometry of molecules reacting by an electron transfer-chemical reaction-electron transfer, or ECE, pathway.

The initial potential (E_i) of the working electrode is chosen so that no electrolysis occurs. At time t = 0, the potential is rapidly pulsed to a potential where the first electron transfer is so fast that its rate does not influence the shape of the charge-time response. Under these conditions, the response Q(t) for the ECE model is of the form:

$$Q(t) = b_2 + b_1[2 - (\text{II}/4b_0t)^{1/2}\,\text{erf}(b_0t)^{1/2}] + b_3t. \quad (2)$$

The $b_0, \ldots, b_3$ are regression parameters. The rate constant for the chemical step is $k = b_0$, and is generally the most desired parameter in this type of analysis. Identities of the other parameters are not relevant to our discussion.

Gram-Schmidt orthogonalization was used to transform Eq. (2) to the form.

$$y_j = B_1h_1(t_j) + B_2h_2(t_j) + B_3h_3(t_j; B_0) \quad (3)$$

where the $h_i(t_j)$ are the orthonormal functions, the B_i are the new set of parameters, and $B_0 = k$. The reader is referred to the original literature[3] for details of the mathematical manipulations.

This procedure defines a new orthogonal parameter space. Here, the most desired parameter is B_0, which is determined by the regression analysis. The other original parameters b_i are computed from the B_i estimated in the regression analysis.

Gram-Schmidt orthogonalization of the ECE reaction model in single potential step chronocoulometry totally removed correlation between parameters when using the Marquardt-Levenberg nonlinear regression algorithm.[3] The orthogonalized ECE model was used to estimate chemical rate constants for decomposition of unstable anion radicals produced during the chronocoulometric reduction of aryl halides in organic solvents. The orthogonalized model eliminated divergence problems and converged three to four times faster than the nonorthogonal

model, making real-time kinetic analyses possible on a microcomputer. Precision on the order of ±6% for $k \leq 10$ s^{-1} was obtained.

II. Secondary Structure of Proteins by FT-IR

Introduction. Proteins are biopolymers consisting of polypeptide chains of amino acid molecules linked in a linear fashion. In biological systems, proteins function as catalysts for life supporting chemical reactions and as structural components of living organisms. In their native state, polypeptide chains fold in a complicated manner which is essential to their biological function. Folding patterns of proteins may be characterized by periodic structures such as helices, sheets, and extended portions. Other structural units include a variety of turns, loops, and disordered coils. Determination of the way the protein is folded is called secondary structural analysis.

Secondary protein structure can be determined by several types of instrumental methods such as x-ray crystallography, nuclear magnetic resonance, circular dichroism, and infrared spectroscopy. Since the development of commercial FT-IR spectrometers, methods for analyzing IR data are being developed to a high degree of accuracy and precision for determination of global secondary structure of proteins.

The backbone of the polypeptide chain absorbs infrared radiation, which excites vibrational modes of chemical linkages called amide bonds. Two of these vibrational modes are of primary importance. The first, the amide I vibration, is primarily caused by stretching of carbon-oxygen double bonds. The amide II vibration is due to stretching of the nitrogen-hydrogen bonds. Infrared spectroscopy measures the amount of light absorbed due to these vibrations over a range of frequencies of the incident light.

The frequencies at which amide I and amide II bands appear are highly dependent on the secondary structure of the protein. However, individual peaks for these vibrational transitions are severely overlapped in FT-IR spectra. This overlap needs to be resolved before a complete structural analysis of the protein can be made. Nonlinear regression analysis coupled with Fourier deconvolution has been successfully applied to this problem.

FT-IR Analysis of Lysosyme. A typical FT-IR spectrum of hen egg white lysosyme showing amide I and amide II regions is given as the outer envelope in Figure 1. This spectrum can be considered a sum of a variety of individual bands which have been assigned to specific structural units of proteins.[4] Identification of all the components of the spectrum by direct nonlinear regression would be a daunting task. To alleviate this dilemma, we first examine the second derivative of the spectrum (insert, Fig. 2) to find the number (n) of component bands and their approximate positions.

The next step in the analysis is to enhance the resolution of the original spectrum via a Fourier deconvolution algorithm developed by Kauppinen et al.[5] Care must be taken to choose the correct values of line width and resolution enhancement factors used by this algorithm so that the FT-IR spectrum is not over- or under-deconvoluted. Under-deconvolution is recognized by the absence of a band indicated by a peak in the second derivative spectrum. Over-deconvolution results in the appearance of large side lobes in the baseline region of the deconvoluted spectrum.[6]

Quantitatively, analysis of the Fourier deconvoluted spectrum by nonlinear regression analysis is also used to help choose the Fourier deconvolution parameters. A model composed of the sum of a series of gaussian peaks is fit to the deconvoluted spectrum (Fig. 2) by nonlinear regression:

$$A = \sum_{j=1}^{n} h_j[\exp\{-(x - x_j)^2/2W_j] \quad (4)$$

where A is absorbance, W_j is peak width, x_j is frequency in cm^{-1}, and h_j is the peak height. W_j, x_j, and h_j for the n peaks are the parameters optimized by the regression analysis. A baseline term is generally not needed after proper background subtraction.

Quantitative criteria to insure correct deconvolution are: 1) correlation of all band assignments with the second derivative peaks; 2) agreement of calculated and experimental baselines; 3) a standard deviation of regression ≤ experimental noise; 4) a successful fit of the model to the original spectrum using fixed frequencies found by fitting the deconvoluted spectrum. In practice, attainment of these criteria may require several cycles of deconvolution

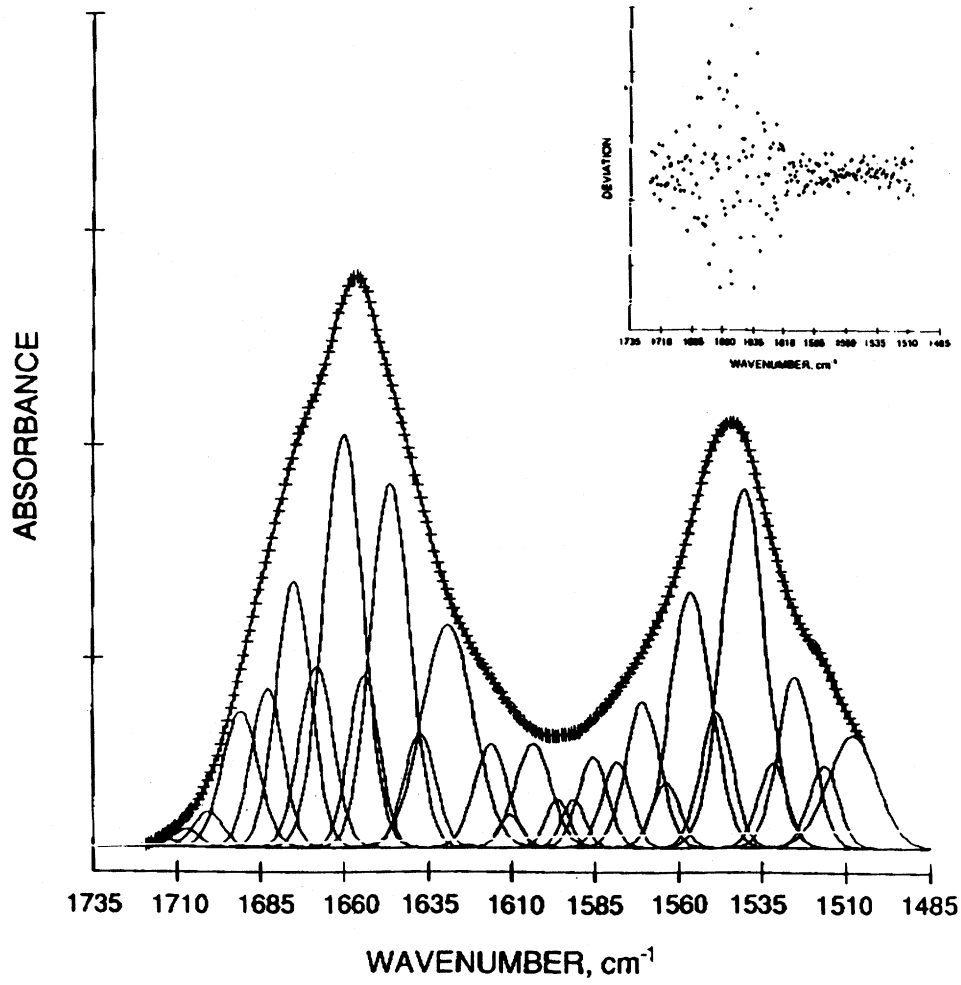

FIGURE 1 FT-IR spectrum showing amide I and amide II bands of lysosyme in aqueous solution. The outer envelope line is the original spectrum. Crosses on the outer envelope and individual component peaks underneath are the results of final regression analysis as described in text. The inset shows plot of residuals or deviations of calculated and experimental absorbances vs frequency.

and regression until optimal deconvolution is achieved.

Criterion 4 involves using the results of the regression analysis of the deconvoluted spectrum (Fig. 2) on Eq. (4) to provide the frequencies of the n bands. These frequencies are fixed, and the model with the same number of bands and frequencies is then fit to the *original spectrum.*

The final fit to the lysosyme FT-IR spectrum is shown in Figure 1 with its 29 component peaks. The inset (Fig. 1) shows that the residuals of the regression are reasonably random, a reliable indication that the model explains the data.[1] Calculated relative areas under the component bands of the original spectrum are in good agreement with those calculated from results of the regression analysis of the fourier deconvoluted spectrum.[6]

Table 1 assigns component bands obtained from the above analysis made by reference to previous vibrational assignments.[4] From relative areas under these bands, we can obtain the fractional amounts of the different structural features in the polypeptide chain. Replicate values of these fractions are obtained from amide I and amide II regions. Good agreement between corresponding fractions adds confidence to this analysis. However, the amide I region gives the best results because the components of the amide II region are inherently more poorly revolved.

The FT-IR structural analysis can be compared with global secondary structure (Fig. 3) determined by x-ray crystallography. Good agreement is obtained (Table 1), noting that the estimated extended and unordered fractions from x-ray analysis are not exact. In this example, fractional amounts of helix and turns from FT-IR and x-ray crystallography are quite comparable. Such comparisons can be used to determine whether the dynamic structure in water, where the protein maintains a structure more relevant to its biological role, is the same as the

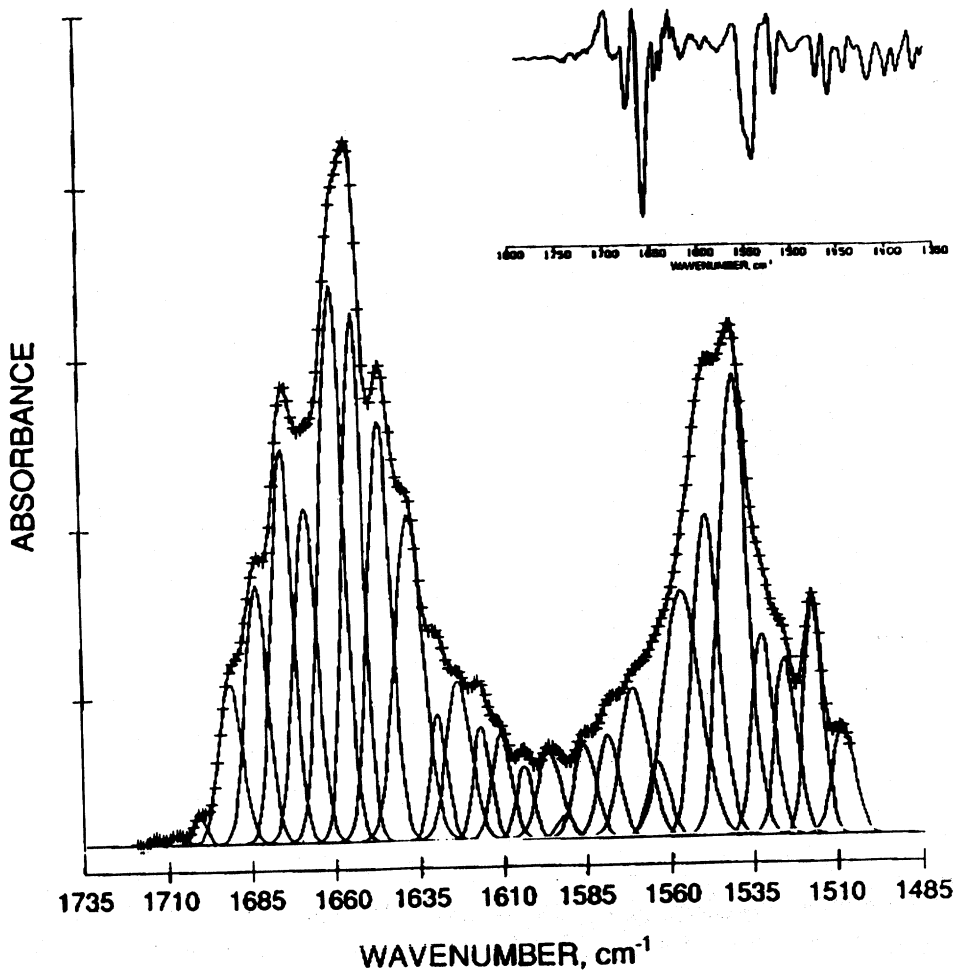

FIGURE 2 Fourier deconvolution of FT-IR spectrum of lysosyme in Figure 1. Crosses on the outer envelope and individual component peaks underneath were found by regression analysis as described in text. The inset shows the second derivative of the original spectrum.

TABLE 1 Secondary structural analysis of FT-IR of lysosyme.

Assignments:	Helix	Extended	Unordered (Loops)	Turns
		(cm^{-1})		
Amide I	1660	1637	1646	1691
	1654	1629		1683
		1623		1675
				1668
Amide II	1540	1532	1548	1578
		1525		1571
				1564
				1556

Fractions:	Helix	Extended	Unordered	Turns
FT-IR Results:		From Deconvoluted Spectrum		
Amide I	0.324 ± 0.014	0.204 ± 0.012	0.131 ± 0.003	0.311 ± 0.013
Amide II	0.278 ± 0.012	0.175 ± 0.010	0.183 ± 0.045	0.364 ± 0.101
		From Original Spectrum		
Amide I	0.260	0.117	0.170	0.318
Amide II	0.323	0.176	0.090	0.411
X-ray Structure	0.310	0.155[a]	0.225[b]	0.310

[a] As fraction of β-turns only, does not include all the extended features of the protein.
[b] Difference between sum of other reported fractions and 1.

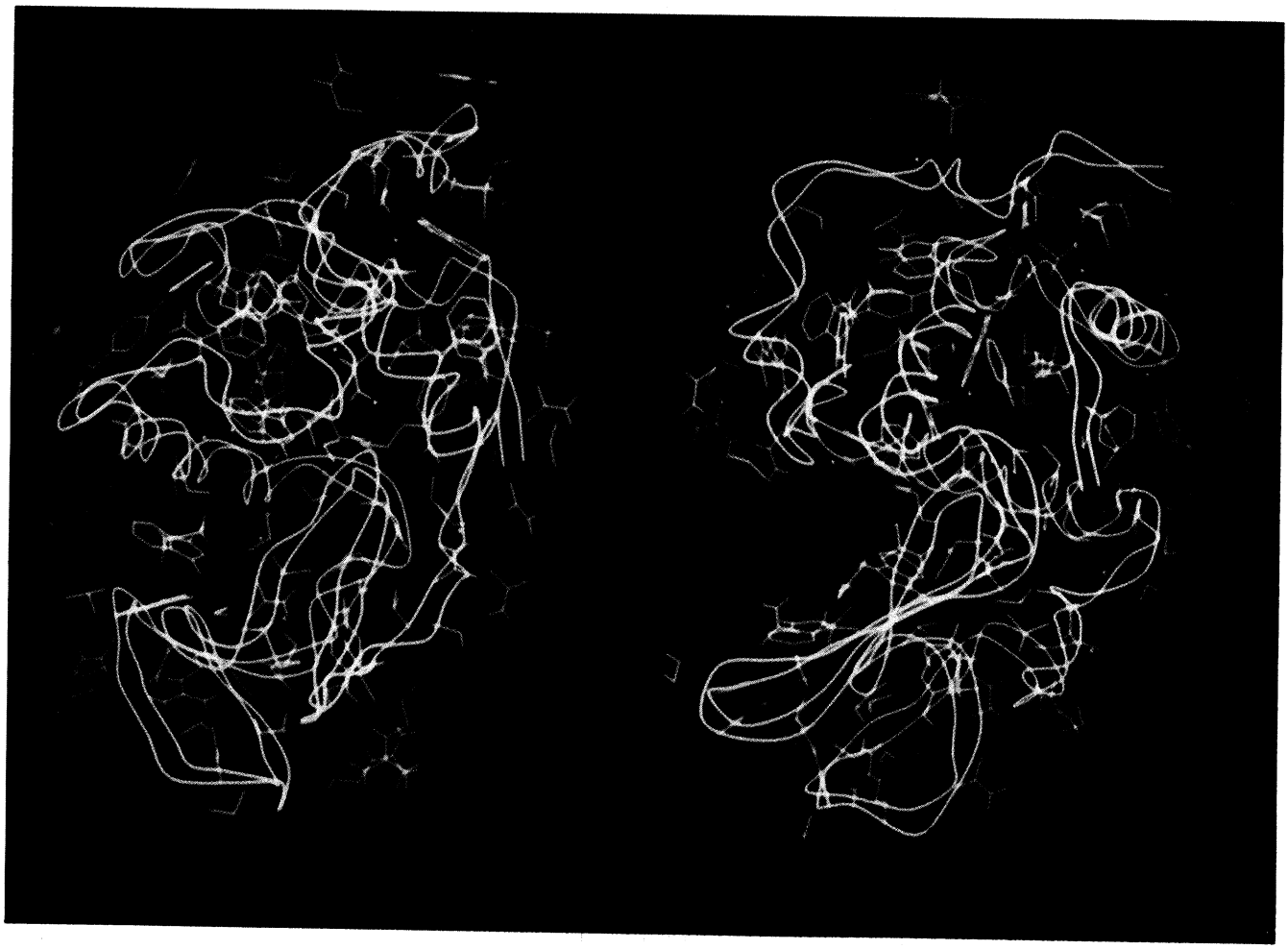

FIGURE 3 Orthogonal representations of the x-ray crystal structure of lysosyme. The right-hand view is rotated 90° from the left-hand view. Side chain color code: green, hydrophobic; red, acidic; purple, basic.

structure of the crystalline protein. The same approach is being applied to other proteins for which crystal structures are available. The ultimate goal of this work is to obtain a statistically significant predictor of secondary structure from the amino acid sequence of the polypeptide.

Conclusions

The above examples demonstrate that in instances where nonlinear regression cannot stand alone, its combination with auxiliary mathematical methods can result in powerful analytical tools. The two methods discussed are general and complimentary. Gram-Schmidt orthogonalization is applicable to highly correlated models with small numbers of parameters. Clearly, the number of mathematical manipulations to orthogonalize the model for the lysosyme FT-IR spectrum is prohibitive; the model ends up having 29 peaks and 87 parameters. However, the combination of nonlinear regression with resolution-enhanced Fourier deconvolution is able to provide a physically meaningful and self-consistent analysis. These two methods should enable extension of nonlinear regression to types of data that were heretofore impossible to analyze with high precision.

This work was supported in part by U.S. PHS Grant No. ES03154 awarded by the National Institutes of Health through the National Institute of Environmental Health Sciences.

References

1. Rusling JF. *CRC Crit. Rev. in Anal. Chem.*, 21, 1989, 49–81.
2. Kumosinski, T.F., in *Advances in Food and Nutrition Research,* Vol. 34, J.F. Kinsella, ed., Academic Press, NY, 1990, pp. 299–385.
3. Sucheta, A. and Rusling, J.F., *J. Phys. Chem.* 93, 1989, 5796–5802.

4. Krimm, S. and Bandekar, J., *Adv. Protein Chem.* 38, 1986, 183–3363.
5. Kauppinen, J.K., Moffatt, D.J., Mantsch, H.H., and Cameron, D.G., *Appl. Spec.* 35, 1981, 271–276.
6. Curley, D., Kumosinski, T.F., Farrell, H.M., and Smuda, E., *Biophys. J.*, 61, 1992, A341.
7. Software for these computations can be obtained by writing to Dr. William Damert, Eastern Regional Research Center, U.S. Department of Agriculture, Wyndmoor, PA 19118.